REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)* MARCH 2009	**2. REPORT TYPE** Conference Paper Preprint	**3. DATES COVERED** *(From - To)* March 2008 – April 2009

4. TITLE AND SUBTITLE
TEST TOKEN DRIVEN ACOUSTIC BALANCING FOR SPARSE ENROLLMENT DATA IN COHORT GMM SPEAKER RECOGNITION (PREPRINT)

5a. CONTRACT NUMBER
FA8750-09-C-0067 & 05-C-0029

5b. GRANT NUMBER
N/A

5c. PROGRAM ELEMENT NUMBER
35885G

6. AUTHOR(S)

Jun-Won Suh, and John H.L. Hansen

5d. PROJECT NUMBER
3188

5e. TASK NUMBER
BA

5f. WORK UNIT NUMBER
AE

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Research Associates for Defense Conversion, Inc.
10002 Hillside Terrace
Marcy, NY 13403-2102

CRSS
University of Texas at Dallas
Richardson, TX 75083

8. PERFORMING ORGANIZATION REPORT NUMBER

N/A

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

AFRL/RIEC
525 Brooks Road
Rome NY 13441-4505

10. SPONSOR/MONITOR'S ACRONYM(S)
N/A

11. SPONSORING/MONITORING AGENCY REPORT NUMBER
AFRL-RI-RS-TP-2010-8

12. DISTRIBUTION AVAILABILITY STATEMENT
Approved for public release; distribution unlimited. PA#: 88ABW-2009-1434, Date Cleared: 08-April-2009

13. SUPPLEMENTARY NOTES
This work, resulting in whole or in part from Department of the Air Force contract number FA8750-05-C-0029, has been submitted to but not accepted for publication at the 2010 International Conference on Acoustics, Speech, and Signal Processing.

14. ABSTRACT
For this study, we address the problem to in-set/out-of-set speaker recognition with sparse enrollment data. Sparse enrollment data presents a unique challenge due to a lack of acoustic space coverage. The proposed algorithm focuses on filling acoustic holes and fortifying the phone expectation in the test stage. This scheme is possible by using the GMM model to classify the speaker phone information at the feature level. The parallel training for most occurred (top) and less occurred (bottom) rank ordered mixture classification (speaker phone class) information is called “Sweet-16”, and the employing a test data mixture histogram using the Sweet-16 is called “Sweet-16 On-The-Fly (OTF)”. The Sweet-16 OTF method is evaluated using telephone conversation speech from the FISHER corpus. The Sweet-16 OTF improves on average 2.17% absolute EER over the previousSweet-16, and average 4.03% absolute EER over GMM-UBM baseline using 2sec test data. The proposed algorithm improvement is a noteworthy stage to compensate for both sparse enrollment data and limited test data.

15. SUBJECT TERMS
In-set/out-of-set speaker recognition, cohort speakers, data sparseness, speaker adaptation, speaker similarity

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON** John G. Parker, Jr.
a. REPORT U	**b. ABSTRACT** U	**c. THIS PAGE** U	UU	5	**19b. TELEPHONE NUMBER (*Include area code)*** N/A

TEST TOKEN DRIVEN ACOUSTIC BALANCING FOR SPARSE ENROLLMENT DATA IN COHORT GMM SPEAKER RECOGNITION

Jun-Won Suh, John H. L. Hansen

Center for Robust Speech Systems (CRSS)
Erik Jonsson School of Engineering & Computer Science
University of Texas at Dallas
Richardson Texas 75083-0688, U.S.A.
{jxs064200, John.Hansen}@utdallas.edu

ABSTRACT

For this study, we address the problem of in-set/out-of-set speaker recognition with sparse enrollment data. Sparse enrollment data presents a unique challenge due to a lack of acoustic space coverage. The proposed algorithm focuses on filling acoustic holes and fortifying the phone expectation in the test stage. This scheme is possible by using the GMM model to classify the speaker phone information at the feature level. The parallel training for most occurred (top) and less occurred (bottom) rank ordered mixture classification (speaker phone class) information is called "Sweet-16", and the employing a test data mixture histogram using the Sweet-16 is called "Sweet-16 On-The-Fly (OTF)". The Sweet-16 OTF method is evaluated using telephone conversation speech from the FISHER corpus. The Sweet-16 OTF improves on average 2.17% absolute EER over the previous Sweet-16, and average 4.03% absolute EER over GMM-UBM baseline using 2sec test data. The proposed algorithm improvement is a noteworthy stage to compensate for both sparse enrollment data and limited test data.

Index Terms— in-set/out-of-set speaker recognition, cohort speakers, data sparseness, speaker adaptation, speaker similarity

1. INTRODUCTION

In-set/out-of-set speaker recognition provides a binary decision for a claimed speaker based on a predefined speaker model from an in-set group. The extended application can be found in identifying speakers in a multi-speaker conversation or broadcast news, or the system grants security access for a specific group in organizations. A speaker's intrinsic and extrinsic traits have previous been studied to achieve robust speaker recognition using clustering[1], discriminative training[2], or high level information[3]. The Gaussian Mixture Model (GMM) provides robust text independent speaker recognition system[4][2]. The statistical model represents the most common characteristics of the available speaker data. The speaker independent model is constructed with a development speaker group to represent out-of-set speaker known as Universal Background Model (UBM). As the out-of-set speaker group becomes larger, the UBM plays a crucial role to decide the speakers identify, such as in the NIST Speaker Recognition Evaluation (SRE) task.

In this study, we focus on sparse enrollment data (5sec) with short test utterances (2∼6 sec) for the in-set/out-of-set problem. The sparse enrollment data results in a unique challenge due to a lack of acoustic phone coverage compared with longer conversational speech data, and the acoustic phone coverage becomes of high risk to evaluate with short test utterances. We called this phenomena the "acoustic hole in the acoustic model space". We focus here to fill the acoustic holes and fortifying the phoneme in the sparse enrollment data to reduce the equal error rate (EER) of system performance. The phone classification is achieved using a Speaker Independent GMM (S.I.GMM), and the classified speaker phone information facilitates the speaker model to fill acoustic holes and to reinforce the phones not seen in the enrollment stage. The proposed system attempts to achieve a major impact by employing a test data phone information distribution. If the test data is shorter than the enrollment data, the proposed algorithm focuses on fortifying the expecting phones in the test stage. The resulting speaker model focuses only on 2sec test data phone information, so the model will generally have better discrimination for 2sec data. For other cases, the longer test data provides further information an phoneme coverage than in the enrollment data. Here, separate training for the top and bottom rank ordered mixture index classification information is called "Sweet-16", and employing the test frame data mixture index histogram labeled using the Sweet-16 is called "Sweet-16 On-The-Fly (OTF)". This approach identifies the acoustic holes with more information to increase the

This project was funded by AFRL under a subcontract to RADC Inc. under FA8750-05-C-0029, and by Univ. of Texas at Dallas under project EMMITT.

probability of filling acoustic holes using a parallel training strategy.

This paper is organized as follows. Sec. 2 explains the baseline system for evaluating the proposed algorithm. Sec. 3 presents motivation and a detailed procedure for developing the proposed algorithm. Next, the evaluation and results of proposed algorithm is accessed with baseline systems in Sec. 4.2. Finally, conclusions and future work is discussed in Sec. 5.

2. BASELINE SYSTEM

2.1. In-set/Out-of-set Speaker Recognition

We assume we are given a set of in-set (enrolled) speakers, and an organized collected data set $\mathbf{X}_n$, corresponding to each enrollment speaker S_n, $1 \leq n \leq N_{in\text{-}set}$. Let the data $\mathbf{X}_0$ represent all outside non-enrolled speakers in the development set. Each speaker dependent statistical model Λ_n, $\{\Lambda_n \in \mathbf{\Lambda}, 1 \leq n \leq N_{in\text{-}set}\}$, can be obtained from $\mathbf{X}_n$. In the first stage, called *(closed-set) speaker identification*, we first classify X into one of the most likely in-set speakers Λ^* as

$$\Lambda^* = \underset{1 \leq n \leq N_{in-set}}{\operatorname{argmax}} \; p(\mathbf{X}|\Lambda_n) \tag{1}$$

In the second stage, called *speaker verification*, we verify whether the observation $\mathbf{X}$ truly belongs to Λ^* or not (i.e., accept/reject).

2.2. GMM-UBM Baseline

The most recognized text-independent system uses Gaussian Mixture Model (GMM) to represent the out-of-set model for outliers (e.g. UBM) and to adapt the speaker into the in-set speaker model with Maximum A Posteriori (MAP)[4][2]. A speaker model is represented by M components of Gaussians trained from the D dimensional observation vector $\mathbf{x}_t$. A GMM is denoted as $\Lambda_n = (\boldsymbol{\omega}_{nm}, \boldsymbol{\mu}_{nm}, \boldsymbol{\Sigma}_{nm})$, for $m = 1, \ldots, M$ and $n = 1, \ldots, N$ where ω_{nm} is the mixture weight of the mth component unimodal Gaussian density $\mathcal{N}_{nm}(\mathbf{x_t})$, with each parameterized by a mean vector $\boldsymbol{\mu}_{nm}$ and covariance matrix $\boldsymbol{\Sigma}_{nm}$, which is assumed diagonal

$$\mathcal{N}_{nm}(\mathbf{x_t}) = \frac{1}{(2\pi)^{\frac{D}{2}} |\boldsymbol{\Sigma}_{nm}|^{\frac{1}{2}}} e^{-\frac{1}{2}(\mathbf{x}_t - \boldsymbol{\mu}_{nm})^T \boldsymbol{\Sigma}_{nm}^{-1} (\mathbf{x}_t - \boldsymbol{\mu}_{nm})}. \tag{2}$$

2.3. GMM-Cohort UBM Baseline

The speaker dependent model is built with MAP using only mean adaptation from UBM in Sec. 2.2, the resulting GMM represents a simple rotation of the same Gaussian mixture densities of the UBM. The acoustic holes caused for sparse in-set data are effectively filled with the Cohort UBM[5]. Since the cohort UBM is built with $N_{cohort}(\ll N_{dev})$ speaker data, the resulting Gaussian mixture density represents a precise acoustic space for speaker phone information versus the UBM. Here, the precise speaker similarity measure improves the overall system[6]. The procedure is as follows:

Step 1: Collect a mixture tagged feature (see Sec. 3.2 for GMT), $\mathbf{X}_n^{mix} = \{\mathbf{x}_n^{mix^1}[p], \mathbf{x}_n^{mix^2}[q], \ldots, \mathbf{x}_n^{mix^m}[r]\}$ (p, q, r arbitrary number of vector $\mathbf{x}_n^{mix}$ element), from the GMT resulting feature,

$$\mathbf{X}_n^{tagged} = \{x_{n1}^{mix^l}, x_{n2}^{mix^l}, \ldots, x_{nT_n}^{mix^l}\} \; 1 \leq l \leq \mathrm{m}$$

for the mth components of GMT, for enrollment speaker n, $1 \leq n \leq N_{in\text{-}set}$. Each mixture represents speaker phone-like information.

Step 2: Collect equal amounts of development feature,

$$\mathbf{X}_i^{mix} = \{\mathbf{x}_i^{mix^1}[p], \mathbf{x}_i^{mix^2}[q], \ldots, \mathbf{x}_i^{mix^m}[r]\} 1 \leq i \leq N_{dev},$$

corresponding to in-set data, $\mathbf{X}_n^{mix}$. Both speaker features should have the same number of mixture classes. Each mixture class for in-set and development should have an equal number of features, $\mathbf{x}_n^{mix^1}[p] = \mathbf{x}_i^{mix^1}[p]$.

Step 3: Build speaker models for both $\Lambda_n^{in\text{-}set}$ and $\mathbf{\Lambda}_n^{dev}$ for each in-set speaker n.

Step 4: Compute the $N_{in\text{-}set} \times N_{dev}$ acoustic space distance matrix between enrollment and development GMMs using the KL divergence, as follows:

$$\begin{aligned} KL(\Lambda_n^{in\text{-}set}, \Lambda_i^{dev}) \quad = \quad & E_{\Lambda_n^{in\text{-}set}(X)}[log \frac{\Lambda_n^{in\text{-}set}(X)}{\Lambda_i^{dev}(X)}] + \\ & E_{\Lambda_n^{dev}(X)}[log \frac{\Lambda_n^{dev}(X)}{\Lambda_i^{in\text{-}set}(X)}] \end{aligned} \tag{3}$$

Step 5: Sort the KL distance score, and pick the top N_{cohort} from a rank ordered development speaker set.

Step 6: Build Λ_n^{cohort} using the top N_{cohort} speaker data.

Step 7: Adapt the speaker model $\Lambda_n^{in\text{-}set}$ from Λ_n^{cohort} with in-set data.

3. PROPOSED ALGORITHM

3.1. Motivation

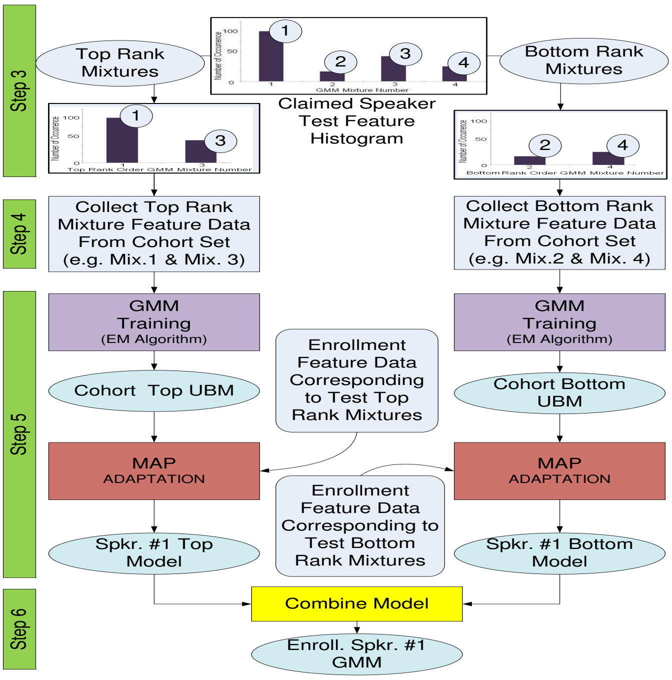
A speaker recognition system with sparse enrollment data will have a difficult time in decoding the legitimacy of speakers identity given extremely short test data 2sec. The acoustic space of a 5sec in-set speaker data is far from what is needed to represent the entire in-set speaker acoustic space. We exploit an acoustically similar speakers phoneme data to fill in for sparse in-set data[5]. A previous proposed system [6] enables us to exploit the specific speaker phone information, and it briefly noted in Sec. 3.2. For exceptionally short test data (2sec), the speaker model should not misrecognize the phones, which have been trained for the enrollment stage. The robust distinction for trained phones would impact system performance. A longer test utterance than training in-set data can take advantage of deciding which phone information is filled or needs to be filled. Since the test data is 2∼6sec, the test data is instantaneously categorized and quantized to each mixture of GMM "on the fly". We assume that each mixture of GMM represents the speaker phone information. Consequently, the emphasis on speaker modeling using test speakers phone distribution information effect the better representation of speaker model for the given test data.

3.2. GMM Mixture Tagging(GMT)

The short amount of data requires exploiting information from acoustically similar speakers. Additionally, the data separation enables us to build a discriminating model for specific targets. The phoneme is one category to parse the speech information. We employ GMM to represent the speaker phone information by each mixture. The GMM is built with developments and in-set speakers data, so we call this the Speaker Independent GMM (S.I. GMM). The speech feature frames are tagged with the highest probability mixture of the S.I. GMM. The test feature frames are also labeled with the S.I.GMM, when claimed speaker provides his/her speech into system.

3.3. Sweet 16 On The Fly

The primary procedure here is similar to that presented with in Sec. 2.3, with the major difference being that a histogram of the mixtures is used to tag feature frame data. The procedure to build the in-set speaker model is as follows:



Step 3
Top Rank Mixtures
Bottom Rank Mixtures
100
50
0
Number of Occurrence
1
2
3
4
GMM Mixture Number
Claimed Speaker Test Feature Histogram
Top Rank Order GMM Mixture Number
Bottom Rank Order GMM Mixture Number
Step 4
Collect Top Rank Mixture Feature Data From Cohort Set (e.g. Mix.1 & Mix. 3)
Collect Bottom Rank Mixture Feature Data From Cohort Set (e.g. Mix.2 & Mix. 4)
Step 5
GMM Training (EM Algorithm)
GMM Training (EM Algorithm)
Enrollment Feature Data Corresponding to Test Top Rank Mixtures
Cohort Top UBM
Cohort Bottom UBM
MAP ADAPTATION
MAP ADAPTATION
Enrollment Feature Data Corresponding to Test Bottom Rank Mixtures
Spkr. #1 Top Model
Spkr. #1 Bottom Model
Step 6
Combine Model
Enroll. Spkr. #1 GMM

Fig. 1. *Block diagram of Sweet 16 OTF. Each step is described in Section 3.3*

Step 1: Select the most acoustically similar speaker set for each in-set speaker n, $1 \leq n \leq N_{in\text{-}set}$.

Step 2: Label the in-set and development speech feature frame data with a 32 mixture class using GMT

Step 3: The process continues by counting the most occurring 16 mixture classes(top 16) and the least occurring 16 mixture classes(bottom 16) for the *claimed speaker's feature*. Make a mixture histogram for the claimed speaker.

Step 4: Pool the top/bottom frame data of the selected cohorts and construct a cohort GMM as $\Lambda_n^{top\text{-}cohort}$ and $\Lambda_n^{bottom\text{-}cohort}$ using the claimed speaker's histogram.

Step 5: Using $\Lambda_n^{top\text{-}cohort}$ and $\Lambda_n^{bottom\text{-}cohort}$ as an initial model for the mean, covariance, and mixture weights, build the in-set speaker model Λ_n^{top} and Λ_n^{bottom} using MAP with the corresponding claimed speaker's histogram.

Step 6: Combine models Λ_n^{top} and Λ_n^{bottom} to build the final in-set speaker model.

4. EXPERIMENTAL RESULTS

4.1. Fisher Corpus

An experiment is performed to evaluate in-set/out-of-set speaker recognition with the telephone conversation corpus, FISHER. The selected 60 speakers are comprised of in-set and out-of-set speakers. We make three different groups of in-set/out-of-set speakers to evaluate group size, 15in/45out, 30in/30out, and 45in/15out. All 60 speakers are devoted to the in-set or out-of-set groups with 50 randomly chosen combinations for three different groups. The development set consists of 378 speakers having 30 sec of speech data. The analysis window size is set to 20 ms with a 10 ms skip rate. Static 19-dimension Mel-Frequency Cepstral Coefficients (MFCC) are extracted and used for statistical modeling. Silence and low-energy speech parts are removed using an energy based detection technique.

4.2. Evaluations

4.2.1. Basline System

The speaker GMM consists of 32 mixtures to represent speaker traits for the short training data. The UBM model will reflect the out-of-set speaker model or outlier, and it is built with 60 randomly selected speakers from among the 378 speaker development set. The remaining 318 speakers are used to represent a potential cohort speaker pool to fill acoustic holes for the in-set speaker, and we note that this 318 speaker set does not overlap with the 60 speakers used for the UBM. The top 5 cohort speakers are selected across all Cohort evaluation based on the UBM system. With these selected cohort speakers, each in-set speaker cohort model is built with 150 sec of data. This cohort model is then adapted with the 5sec in-set training data via the MAP algorithm.

Sweet-16 is first introduced in a previous study[6], and the present On-The-Fly (Sweet-16 OTF) training method was presented in Sec. 3.3. The primary difference is that the 5sec training data histogram is used to rank the mixture tagged data, as opposed to using the test data histogram. Table 1 shows that the Sweet-16 OTF improves in-set speaker recognition EER by an average 2.17% absolute over the Sweet-16, and an average 4.03% absolute EER over the GMM-UBM Baseline system using only 2sec of test data.

Table 1. EER(%) performance comparison using 2sec test data.

	EER		
	15in/45out	30in/30out	45in/15out
GMM-UBM Baseline	30.62	31.27	31.55
GMM-Cohort UBM Baseline	32.96	32.10	30.43
Sweet-16	26.71	29.13	32.02
Sweet-16 OTF	25.27	26.77	29.30

4.2.2. *Sweet-16 OTF*

The proposed Sweet-16 OTF algorithm employs a cohort speaker group of 5 speakers, the same size which is used for the GMM-Cohort Baseline system. The combined weight ratio is set to 7:3 for the top and bottom GMM speaker model. The resulting mixture weights of the GMM will not sum up to 1 because of the blending of the two models, so this issue needs to be addressed in future work. By employing the mixture tagged test data histogram, the system improves EER on average 2.34% over Sweet-16 on 2 and 6 sec test data. Fig. 2 shows that the equal error rate is reduced by between 2.2%∼6.49% absolute value over the GMM-UBM Baseline. Fig. 2 also indicates that a smaller in-set group tends to produce a lower equal error rate. The large in-set group increases the distinction perplexity between in-set and out-of-set models, and therefore we expect a higher EER for large in-set groups. In summary, the proposed method impacts system performance by focusing the expected phone information data and harvesting unseen phone information collected from feature frame level data.

5. CONCLUSIONS AND FUTURE WORKS

In this study, we have developed a novel strategy to enforce an improved data training balance for the speaker model using the expected phone information from a test data mixture tagged histogram for 2sec test data. The Sweet-16 strategy improves acoustic hole filling, resulting from the limited in-set speaker data. Evaluations were performed the "landline telephone channel" from FISHER corpus to avoid handset variation, and focus on acoustic hole filling. The proposed Sweet-16 OTF training method improves in-set speaker recognition EER by 2.2∼6.49% absolute with 2∼6sec of test data. Future work could consider expecting the method to normalize for handset variation effect with the FISHER corpus so that cohort speakers can be selected from any corpus.

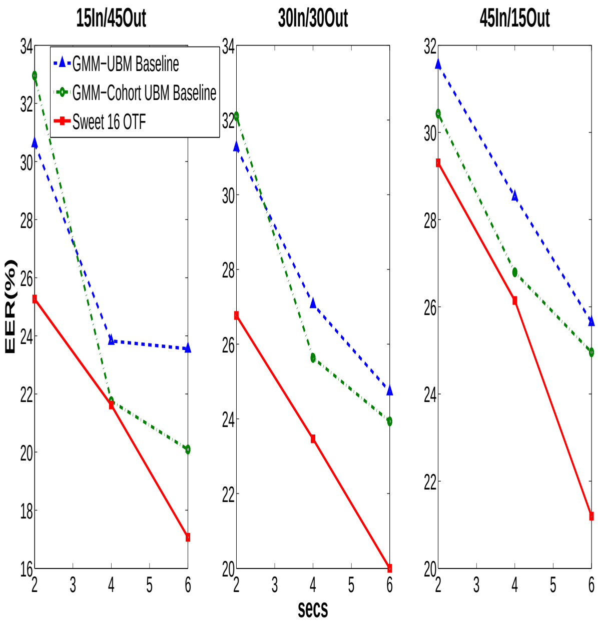


Fig. 2. *Performance (in terms of EER(%)) of baseline and proposed algorithm on FISHER, using in-set/out-of-set speaker sizes of 15/45, 30/30 and 45/15.*

6. REFERENCES

[1] R. Kuhn, J.-C. Junqua, P. Nguyen, and N. Niedzielski, "Rapid speaker adaptation in eigenvoice space," in *IEEE Trans. Audio, Speech, & Language Proc.*, Nov. 2000, vol. 8, pp. 695–707.

[2] D.A. Reynolds, T.F. Quatieri, and R.B. Dunn, "Speaker verification using adapted gaussian mixutre models," in *Digital Signal Proc.*, 2000, vol. 10, pp. 19–41.

[3] W.M. Campbell, J.R. Campbell, D.A. Reynolds, and T.R. Leek Jones, "High-level speaker verification with support vector machines," in *Acoustics, Speech, and Signal Processing, 2004. Proceedings. (ICASSP '04)*, May 2004, vol. 1, pp. 73–76.

[4] H. Gish and M. Schmidt, "Text-independent speaker identification," in *IEEE Signal Process. Mag.*, Oct. 1994, vol. 11, pp. 18–32.

[5] V. Prakash and J.H.L. Hansen, "In-set/out-of-set speaker recognition under sparse enrollment," in *IEEE Trans. Audio, Speech, & Language Proc.*, Sep. 2007, vol. 15, pp. 2044–2052.

[6] J.-W Suh, P. Angkititrakul, and J.H.L. Hansen, "Filling acoustic holes through leveraged uncorellated gmms for in-set/out-of-set speaker recognition," in *Interspeech Conf. 2008*, 2008.